

A system and method for controlling command queuing on parity drives in an array of disk drives.

Patent Number: ☐ EP0661635, B1
Publication date: 1995-07-05
Inventor(s): ISLAM SHAH MOHAMMAD REZAUL (US)
Applicant(s):: IBM (US)
Requested Patent: ☐ JP7210334
Application Number: EP19940309362 19941214
Priority Number(s): US19930175710 19931230
IPC Classification: G06F11/10 ; G06F12/08
EC Classification: G06F11/10M, G06F12/08B12
Equivalents: DE69412775D, DE69412775T, KR162124, ☐ US5530948

Abstract

The present invention provides a technique for queuing commands on an input/output controller for a parity drive in a level 4 or level 5 redundant array of inexpensive disk drives, which responds to receipt of a write instruction with appended data by determining a logical block address of a data drive for the appended data and a logical block address of a parity drive for redundant data belonging to a destination stripe of the appended data. If a parity cache entry corresponds to the parity address, a cache hit is deemed to have occurred. Responsive to a cache hit, computation of replacement parity data for the stripe may be done from parity data in cache at a location given by the parity cache entry, data read from the data drive, and the appended data. The newly calculated parity data is then placed in cache. A command is then sent to an input/output controller for the drive where the parity data is located to write the replacement parity data. Serialization of access to the parity data is moved from the disk drive to cache memory, allowing command

queuing to be used with the disk drive to preserve optimum available performance of the drive.



Data supplied from the esp@cenet database - I2

(19) 日本国特許庁 (J P)

(12) 公開特許公報 (A)

(11) 特許出願公開番号

特開平7-210334

(43) 公開日 平成7年 (1995) 8月11日

(51) Int. Cl. ⁶	識別記号	序内整理番号	F I	技術表示箇所
G 0 6 F 3/06	5 4 0			
	3 0 5 C			
11/10	3 2 0 E			
12/16	3 2 0 L	9293-5B		

審査請求 有 請求項の数29 OL (全 14 頁)

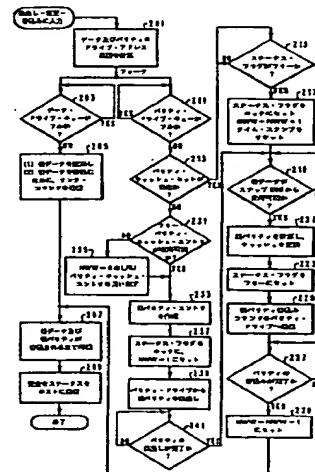
(21) 出願番号	特願平6-236301	(71) 出願人	390009531 インターナショナル・ビジネス・マシーンズ・コーポレーション INTERNATIONAL BUSINESS MACHINES CORPORATION アメリカ合衆国10504、ニューヨーク州アーモンク (番地なし)
(22) 出願日	平成6年 (1994) 9月30日	(72) 発明者	シャ・モハメド・レザウル・イスラム アメリカ合衆国33431、フロリダ州ボカ・ラトン、アパートメント ジェイ226、ジャイウッド・テラス 3380
(31) 優先権主張番号	1 7 5 7 1 0	(74) 代理人	弁理士 合田 潔 (外2名)
(32) 優先日	1993年12月30日		
(33) 優先権主張国	米国 (US)		

(54) 【発明の名称】 データ記憶方法及びキューイング方法

(57) 【要約】

【目的】 集合的に1個以上の論理大容量記憶装置として機能する大容量記憶装置のセットを提供する。

【構成】 RAIDレベル4または5のパリティ・ドライブの入出力制御装置上におけるコマンド・キューイングが、付加データを伴う書き込み命令の受信にตอบสนองして、データ・ドライブの論理ブロック・アドレスと、パリティ・ドライブの論理ブロック・アドレスを決定する。パリティ・キャッシュ・エントリがパリティ・アドレスに一致すると、キャッシュ・ヒットが発生する。キャッシュ・ヒットにตอบสนองして置換パリティ・データが、パリティ・キャッシュ・エントリにより提供されるキャッシュ位置に存在するパリティ・データ、データ・ドライブから読出されるデータ、及び付加データから計算される。新たに計算されたパリティ・データがキャッシュに配置される。次に置換パリティ・データを書込むために、パリティ・データが配置されるドライブの入出力制御装置にコマンドが送信される。



【特許請求の範囲】

【請求項1】耐障害論理大容量記憶装置内の物理記憶装置にユーザ・データ及び冗長データを記憶する方法であって、

ホストからの付加データを伴う書込みコマンドの受信に
1 応答して、第1の物理記憶装置上における付加データの
論理ブロック・アドレスと、第2の物理記憶装置上にお
ける冗長データのバリティ・アドレスとを決定するステ
ップと、

前記第1の記憶装置に対して、論理ブロック・アドレス
の旧データを読み出し、論理ブロック・アドレスに付加デ
ータを書込むように指令するステップと、

前記第2の記憶装置上の冗長データがメモリ・バッファ
内に存在するかどうかを判断するステップと、

冗長データが前記メモリ・バッファ内に存在しない場
合、前記第2の記憶装置から前記メモリ・バッファに冗
長データを読み出すステップと、

冗長データ、付加データ及び旧データから新たな冗長デ
ータを計算するステップと、

前記第2の物理記憶装置に新たな冗長データを論理プロ
ック・アドレスに記憶するように指令するステップと、
を含む記憶方法。

【請求項2】新たな冗長データを計算するステップが、
前記メモリ・バッファ内の冗長データを新たな冗長デー
タにより置換するステップを含む、請求項1記載の記憶
方法。

【請求項3】ディスク・ドライブの冗長アレイ内のバリ
ティ・ドライブ上にコマンドをキューイングする方法で
あって、

ホストからの付加データを伴う書込みコマンドの受信に
2 応答して、データ・ドライブ上における付加データの論
理ブロック・アドレスと、前記バリティ・ドライブ上にお
けるバリティ・データの論理ブロック・アドレスとを
決定するステップと、

バリティ・キャッシュ・エントリがバリティ・データの
論理ブロック・アドレスに一致するかどうかを判断する
ステップと、

前記一致の肯定判定に
3 応答して、バリティ・キャッシュ・エントリにより提供されるバッファ・ロケーションの
バリティ・データと、前記データ・ドライブから読出さ
れるデータと付加データとから、置換バリティ・データ
を計算するステップと、

置換バリティ・データを前記バッファ・ロケーションに
配置するステップと、

置換バリティ・データを書込むためのコマンドを、バリ
ティ・データの論理ブロック・アドレスに対応する入出
力制御装置に送信するステップと、

を含むキューイング方法。

【請求項4】前記一致の否定判定に
4 応答して、バリティ・データを前記バッファ・ロケーションに読出するための

コマンドを、前記バリティ・ドライブの論理ブロック・
アドレスに対応する入出力制御装置に送信するステップ
と、

前記バリティ・ドライブの論理ブロック・アドレスのバ
リティ・キャッシュ・エントリを生成するステップと、
を含む、請求項3記載のキューイング方法。

【請求項5】バリティ・キャッシュ・エントリが前記バ
リティ・ドライブの論理ブロック・アドレスと、ドライ
ブ識別と、タイム・スタンプと、ロック／非ロック・フ
ラグと、待機中の書込みの数を示す書込み待機カウンタ
と、メモリ・バッファ内のバリティ・データの位置を示
す位置フィールドとを含む、請求項4記載のキューイン
グ方法。

【請求項6】前記データ・ドライブの論理ブロック・ア
ドレスの決定に
5 応答して、データを読み出し次に付加デー
タを書込むためのリンク・コマンドを、論理ブロック・
アドレスをカバーする入出力制御装置に送信するステッ
プ、

を含む、請求項5記載のキューイング方法。

【請求項7】計算ステップがリンク・コマンドの送信ス
テップの読出しオペレーションの後に発生する、請求項
6記載のキューイング方法。

【請求項8】前記バリティ・キャッシュ・エントリの生
成ステップが、

バリティ・キャッシュ・テーブルが未使用のバリティ・
キャッシュ・エントリを有するかどうかを判断するステ
ップと、

肯定判定に
6 応答して、未使用のバリティ・キャッシュ・
エントリを使用するステップと、

否定判定に
7 応答して、待機中の書込みを有さないバリティ・
キャッシュ・エントリの中から、最低使用頻度のバ
リティ・キャッシュ・エントリを探し出し、探し出され
たバリティ・キャッシュ・エントリを使用するステップ
と、

を含む、請求項5記載のキューイング方法。

【請求項9】置換バリティ・データ書込みのためのコマ
ンドの送信ステップが、書込み待機カウンタを増分する
ステップを含む、請求項5記載のキューイング方法。

【請求項10】置換バリティ・データの生成のために、
8 前記バッファ・ロケーション内のバリティ・データへの
アクセスを直列化するステップ、

を含む、請求項9記載のキューイング方法。

【請求項11】置換バリティ・データが前記バリティ・
ドライブの論理ブロック・アドレスに書込まれた後、バ
リティ・キャッシュ内の書込み待機カウンタを減分する
ステップ、

を含む、請求項10記載のキューイング方法。

【請求項12】前記バリティ・ドライブがチェック・デ
ィスクである、請求項11記載のキューイング方法。

【請求項13】前記バリティ・ドライブがディスク・ド
9

ライブの冗長アレいの任意の1つに相当する、請求項1記載のキューイング方法。

【請求項14】ホスト・データ処理システムからの付加データを伴う書き込み命令の受信に应答して、データ・ドライブ上における付加データの論理ブロック・アドレスと、パリティ・ドライブ上における冗長データの論理ブロック・アドレスとを決定する手段と、

複数のパリティ・キャッシュ・エントリを有するパリティ・キャッシュ・テーブルと、パリティ・キャッシュ・エントリ及び前記パリティ・ドライブ上の論理ブロック・アドレスに対応するパリティ・データを記憶するバッファと、

パリティ・キャッシュ・エントリが発見手段により見い出される前記パリティ・ドライブの論理ブロック・アドレスに一致するかどうかを判断する手段と、

前記一致の肯定判定に应答して、パリティ・キャッシュ・エントリにより提供されるバッファ・ロケーションのパリティ・データと、前記データ・ドライブの論理ブロック・アドレスから読出されるデータと、付加データとから、置換パリティ・データを計算する手段と、置換パリティ・データを前記バッファ・ロケーションに配置する手段と、

置換パリティ・データを書込むためのコマンドを、パリティ・ドライブの論理ブロック・アドレスに対応する入出力制御装置に送信する手段と、

を含む、ディスク・ドライブの冗長アレい。

【請求項15】前記一致の否定判定に应答して、パリティ・データを前記バッファ・ロケーションに読出するためのコマンドを前記パリティ・ドライブの論理ブロック・アドレスに対応する入出力制御装置に送信する手段と、前記パリティ・ドライブの論理ブロック・アドレスから読出されるパリティ・データに対応するパリティ・キャッシュ・エントリを生成する手段と、

を含む、請求項14記載のディスク・ドライブの冗長アレい。

【請求項16】パリティ・キャッシュ・エントリが前記パリティ・データの論理ブロック・アドレスと、タイム・スタンプと、ロック／非ロック・フラグと、待機中の書き込みの数を示す書き込み待機カウンタと、ドライブ識別と、メモリ・バッファ内のパリティ・データの位置フィールドとを含む、請求項15記載のディスク・ドライブの冗長アレい。

【請求項17】前記データ・ドライブの論理ブロック・アドレスの決定に应答して、データを読出し次に付加データを書込むためのリンク・コマンドを、前記データ・ドライブの論理ブロック・アドレスに対応する入出力制御装置に送信する手段、

を含む、請求項16記載のディスク・ドライブの冗長アレい。

【請求項18】パリティ・キャッシュ・エントリの生成

手段が、

パリティ・キャッシュ・テーブルが未使用のパリティ・キャッシュ・エントリを有するかどうかを判断する手段と、

肯定判定に应答して、未使用のパリティ・キャッシュ・エントリを使用する手段と、

否定判定に应答して、待機中の書き込みを有さない最低使用頻度のパリティ・キャッシュ・エントリを探し出し、探し出されたパリティ・キャッシュ・エントリを使用する手段と、

を含む、請求項16記載のディスク・ドライブの冗長アレい。

【請求項19】置換パリティ・データ書き込みのためのコマンドの送信手段が、書き込み待機カウンタを増分する手段を含む、請求項16記載のディスク・ドライブの冗長アレい。

【請求項20】置換パリティ・データの生成のために、前記バッファ・ロケーション内のパリティ・データへのアクセスを直列化する手段、

20 を含む、請求項19記載のディスク・ドライブの冗長アレい。

【請求項21】置換パリティ・データが前記パリティ・ドライブの論理ブロック・アドレスに書込まれた後、前記パリティ・ドライブの論理ブロック・アドレスに対応するパリティ・キャッシュ内の書き込み待機カウンタを減分する手段、

を含む、請求項20記載のディスク・ドライブの冗長アレい。

30 【請求項22】前記パリティ・ドライブがチェック・ディスクである、請求項21記載のディスク・ドライブの冗長アレい。

【請求項23】前記パリティ・ドライブがディスク・ドライブの冗長アレいの任意の1つに相当する、請求項21記載のディスク・ドライブの冗長アレい。

【請求項24】論理装置として動作する物理記憶装置のアレイと、

物理記憶装置のアレイ上の所定のロケーションに記憶される付加データを伴う更新コマンドを提供するコマンド発生器と、

40 ユーザ・データのストライプに対応する選択パリティ・データ・ブロックのコピーを記憶するキャッシュと、ユーザ・データ・ブロック及びユーザ・データ・ブロックのストライプに渡って生成されるパリティ・データ・ブロックを物理記憶装置のアレイ間でストライプし、物理記憶装置ロケーションによりキャッシュ内に記憶される選択パリティ・データ・ブロックの1個以上のコピーを識別するパリティ・キャッシュ・テーブルを保守するローカル・プロセッサと、

50 ローカル・プロセッサによりそれぞれの入出力制御装置に送信されるコマンドをキューイングする、物理記憶装

置の物理アレイの各々に関連する入出力制御装置と、を含むデータ処理システム。

【請求項25】パリティ・キャッシュ・エントリが物理記憶装置の論理ブロック・アドレスと、物理記憶装置の識別と、タイム・スタンプと、ロック／非ロック・フラグと、待機中の書込みの数を示す書込み待機カウンタと、キャッシュ内のパリティ・データ・ブロックの位置フィールドとを含む、請求項24記載のデータ処理システム。

【請求項26】ローカル・プロセッサが、キャッシュ内の選択パリティ・データ・ブロックの各コピーへのアクセスを制御するシリアルライザ、を含む、請求項25記載のデータ処理システム。

【請求項27】キャッシュ内にパリティ・データ・ブロックが存在しないストライプ内の物理アドレスを更新するコマンドにตอบสนองして、前記ストライプに対応するパリティ・データ・ブロックをキャッシュに読出し、パリティ・キャッシュ・テーブルに前記パリティ・データ・ブロックのエントリを生成するパリティ・キャッシュ・テーブル・エントリ発生器、を含む、請求項26記載のデータ処理システム。

【請求項28】パリティ・データが単一の物理記憶装置上に配置される、請求項27記載のデータ処理システム。

【請求項29】パリティ・データが全ての物理記憶装置上に配置される、請求項27記載のデータ処理システム。

【発明の詳細な説明】

【0001】

【産業上の利用分野】本発明は集合的に1個以上の論理大容量記憶装置として機能する大容量記憶装置のセットに関し、特に、RAIDレベル4及び5システムにおけるパリティ・ドライブ上で、コマンドをキューに待機させる（以降ではコマンド・キューイングと記す）システム及び方法に関する。

【0002】

【従来の技術】ディスク・メモリの使用はコンピュータにおいて重要であり続ける。なぜなら、これは不揮発性であり、メモリ・サイズ要求が主メモリの実際の容量を上回り続けるからである。ディスクは主メモリよりも低速であるので、システム性能はしばしばディスク・アクセス速度により制限される。従って、全体的システム性能にとって、メモリ・サイズ及びディスク・ドライブ・ユニットのデータ・アクセス速度を増大することが重要である。この議論に関しては、Michelle Y. Kimによる“Synchronized Disk Interleaving” (IEEE Transactions On Computers, Vol. C-35, No. 11, 1986年11月)を参照されたい。

【0003】ディスク・メモリ・サイズはディスクの数を増やしたり、ディスクの直径を大きくすることにより

増大するが、データ・アクセス速度は向上しない。メモリ・サイズ及びデータ転送レートの両方が、データ記憶密度を増加することにより増大される。データ転送レートはディスクの回転速度を増すことにより向上する。しかしながら、技術的制約がデータ密度を制限し、高密度且つ高速ディスクはエラーを生じ易い。

【0004】データ・アクセス速度を改良するために、様々な技術が使用されてきた。単一トラック上のデータへの連続アクセスにおけるシーク及び回転遅延を排除するために、データの全トラックを保持可能なディスク・キャッシュ・メモリが使用されてきた。ディスク・セット上のまたは単一ディスクのトラック・セット上のデータ・ブロックをインタリーブするために、複数の読出し／書込みヘッドが使用されてきた。一般的データ・ブロック・サイズはバイト・サイズ、ワード・サイズ及びセクタ・サイズである。ディスク・インタリービングは性能を向上させるための既知のスーパーコンピュータ技術であり、前記記事などにおいて述べられている。

【0005】データ・アクセス性能は関連出願に述べられるように、多数のパラメータにより測定される。ランザクション処理（バンキングなど）では、データ転送は、通常、小さく、要求レートは高速且つランダムである。一方、スーパーコンピュータの応用では、大きなデータ・ブロックの転送が一般的である。

【0006】最近開発された比較的低コストで性能が改良されたインタリーブ式ディスク・メモリ・アーキテクチャは、RAID (Redundant Arrays of Inexpensive Disk) としてグループ化される。例えばDavid A. Pattersonらによる“A Case for Redundant Arrays of Inexpensive Disks (RAID)” (Report No. UCB/CSD 87/89, December, 1987, Computer Science Division (EECS), University of California, Berkeley, California 94720)を参照されたい。Pattersonらによる参考文献において述べられるように、大きなパーソナル・コンピュータ市場が、単一大型高価ディスク (SLED: Single Large Expensive Disk) システムに勝る性能対コスト比を有する安価な (inexpensive) ディスク・ドライブの開発を支えてきた。安価なディスクにおける1読出し／書込みヘッド当たりの1秒当たりの入出力数は、大型ディスクの2倍以内である。従って、安価ディスクのセットが単一の論理ディスク・ドライブとして機能するRAIDアーキテクチャにおける、幾つかの安価ディスクからの並列転送は、低価格において、SLEDに勝る性能を提供する。

【0007】しかしながら、データが複数のディスク上に記憶される時、平均故障寿命 (MTTF: mean time to failure) はアレイ内のディスクの数の増加に相反して減少する。システムのこの減少する平均故障寿命を修正するために、誤り検出及び訂正が全てのRAIDアーキテクチャの特徴となる。Pattersonらの参考文献は、

各々が誤り検出及び訂正のための異なる手段を有する5つのRAIDアーキテクチャについて述べている。これらのRAIDアーキテクチャは、RAIDレベル1乃至5として参照される。

【0008】RAIDレベル1はデータの完全な複製("ミラーリング(mirroring)")とも呼ばれる)を使用し、1ディスク当たり比較的低い性能比率を有する。RAIDレベル2は、エラー訂正及びディスク故障回復を提供するために必要な余分なディスク数を低減するエラー訂正コードを使用することにより、1ディスク当たりの容量比率に加え、この性能を改良する。RAIDレベル2では、データはG個のデータ・ディスクのグループにインタリーブされ、単一のエラーを検出し訂正するために、エラー訂正コード(ECC)が生成され、これが"チェック・ディスク"として参照されるC個の追加のディスクのセットに記憶される。ECCはデータ内のランダムな単一ビット・エラーを検出し訂正するために使用され、またG個のデータ・ディスクの1個がクラッシュする場合のデータの回復を可能とする。C+G個のディスクの内のG個だけがユーザ・データを転送するので、1ディスク当たりの性能は $G/(G+C)$ に比例する。 G/C は、通常、1よりも相当に大きく、従ってRAIDレベル2はRAIDレベル1よりも1ディスク当たりの性能を改良する。1個以上のスベア・ディスクがシステム内に含まれ、ディスク・ドライブの1個が故障すると、このスベア・ディスクが電子的にRAID内にスイッチされ、故障したディスク・ドライブを置換することも可能である。

【0009】RAIDレベル3はRAIDレベル2の変形であり、チェック・ディスク数を1個に減らすために、ほとんどの既存の安価ディスク・ドライブにより提供されるエラー検出機能が使用され、それによりRAIDレベル2に比較して、1ディスク当たりの相対性能を向上させる。通常、パリティ・データがECCの代わりに代用される。ECCまたは他のエラー・コード、或いはパリティ・データは冗長データと称される。RAIDレベル2及び3の両方において、大きなデータまたはグループ化データに対するディスク・アクセスのランザクション時間は低減される。なぜなら、全てのデータ・ディスクに対する帯域幅が利用されるからである。

【0010】例えばランザクション処理において一般的である小さなデータ転送の性能基準は、RAIDレベル1乃至3に対して不十分であることが知られている。なぜなら、データがディスク間でビット・サイズまたはバイト・サイズのブロックにインタリーブされ、1セクタより少ないデータへのアクセスにおいてさえも、全てのディスクがアクセスされなければならないからである。この性能パラメータを改良するために、RAIDレベル3の変形であるRAIDレベル4では、レベル1乃至3におけるビットまたはバイト・インタリーブ・モー

ドの代わりに、データがセクタ・インタリーブ・モードによりディスクにインタリーブされる。換言すると、個々の入出力転送が単一のデータ・ディスクだけに関連する。これによる利点は、入出力オペレーションの並列性の潜在に由来する。これは同一のデータ・ディスクを同時にアクセスする別々のデータ・アクセス要求の間の競合を低減する。

【0011】それにも関わらず、RAIDレベル4の性能は、書き込みオペレーションの間のチェック・ディスクに対するアクセス競合のために制限される。全ての書き込みオペレーションにおいて、データが書込まれる各ストライプ(すなわちセクタの行)に対応して、チェック・ディスク上に更新パリティ・データを記憶するために、チェック・ディスクがアクセスされなければならない。Patterson らは、RAIDレベル4及び5において、単一のセクタに対する個々の書き込みが、論理大容量記憶装置内の全てのディスクに関連しないことを観測した。なぜなら、チェック・ディスク上のパリティ・ビットが単にグループ内の全ての対応データ・ビットの単一の排他的論理和に相当するからである。RAIDレベル4では、書き込みオペレーションは常にパリティ・ディスクの読み出し及び再書き込みを含み、パリティ・ディスクが低カレント書き込みオペレーションにおけるアレイへのアクセスのボトルネックとなる。RAIDレベル4の変形であるRAIDレベル5は、パリティ・チェック・データとユーザ・データを全てのディスクに渡って分散することにより、書き込みオペレーションの競合問題を軽減する。RAIDレベル4では、大きな書き込みオペレーション(全てのパリティ・ストライプ・ユニットに広がる)は予備読み出しを要求しない。

【0012】しかしながら、競合問題は依然として発生する。RAIDレベル4及び5は共に、各読み出し・変更・書き込みオペレーション(例えばレコードの更新)において、2個の各ディスクに2回のアクセスを要求する。更新はデータ・ディスク上の既存のユーザ・データの読み出しと、ユーザ・データが帰属するストライプに対応するパリティ・ディスク上のパリティ・データの読み出しを含む。これに続き、更新されたユーザ・データ及びパリティ・データが、それぞれ両方のディスクに書込まれる。読み出しオペレーションは更新パリティを計算するのに不可欠であり、次の関数を用いて実行される。

新パリティ = (旧データ XOR 新データ) XOR 旧パリティ

【0013】パリティ・データのコヒーレンシの喪失を阻止するために、RAIDレベル4及び5の大容量記憶システムにおけるデータ更新オペレーション処理は、原子的または直列化読み出し変更書き込みオペレーションを要求し、その間にパリティ・データを記憶するドライブがロックされ、最初の更新が完了する以前に、別のデータ更新オペレーションのパリティ情報が変化するのを阻止

する。パリティのコヒーレンシとは、パリティ・グループのデータに対して順次的に実行される一連の排他的論理和演算を、パリティが継続的に表すことを意味する。ドライブのロックは、タグド・コマンド・キューイング(TCQ: Tagged Command Queuing)をサポートするディスク・サブシステムにおいて、コマンドのキューイングを阻止する。

【0014】タグド・コマンド・キューイングはSCSI (Small Computer Systems Interface) の規格において定義される。これは応答を待機することなく、ホストによりドライブに送信される複数のコマンドを処理する。コマンド及び応答は、ホストが応答を要求にマッチさせるようにタグ付けされる。幾つかのシステムでは、ドライブ性能を改良するためにオペレーションの実行順序の最適化が行われる。必要に応じて、所定の順序によるコマンドの実行を保証するために、リンク化コマンドが提供される。ドライブに対するアクセスの直列化は、コマンド・キューイングを阻止するので、その後、ディスク・サブシステム制御装置はオペレーション・シーケンスを最適化できず、ディスク・サブシステムの性能に大きく影響を及ぼす結果となる。

【0015】用語"ストライピング (striping)"がRAID技術においてしばしば見受けられる。ストライピングは"ストライプ・ユニット"により、複数のディスク・ドライブに渡ってデータをインタリーブすることである。ストライプ・ユニットは論理的に連続なデータのグループであり、これはデータを異なるディスク上に配置する以前に、単一のディスク上に物理的に連続に書込まれる。

【0016】

【発明が解決しようとする課題】本発明の目的は、集合的に1個以上の論理大容量記憶装置として機能する大容量記憶装置のセットを提供することである。

【0017】本発明の別の目的は、RAIDレベル4及び5システムにおけるパリティ・ドライブ上で、コマンドをキューイングするシステム及び方法を提供することである。

【0018】

【課題を解決するための手段】本発明の上述及び他の目的が次のようにして達成される。ディスク・ドライブのレベル4またはレベル5冗長アレイ内のパリティ・ドライブの入出力制御装置にコマンドをキューイングする方法が、付加データを伴う書込み命令の受信に応答して、付加データに対応するディスク・ドライブ上の論理ブロック・アドレスと、付加データの宛先ストライプに帰属する冗長データに対応するディスク・ドライブ上の論理ブロック・アドレスを決定する。パリティ・キャッシュ・エントリがパリティ・データ論理ブロック・アドレスに一致すると、キャッシュ・ヒットが発生する。キャッシュ・ヒットに反応して、ストライプに対応する置換パ

リティ・データがパリティ・キャッシュ・エントリにより提供されるキャッシュ・ロケーションに存在するパリティ・データ、データ・ドライブから読出されるデータ及び付加データから計算される。新たに計算されたパリティ・データが次にキャッシュに配置される。次に置換パリティ・データを書込むために、パリティ・データが配置されるドライブの入出力制御装置にコマンドが送信される。パリティ・データに対するアクセスの直列化が、ディスク・ドライブからキャッシュ・メモリに転送され、それによりパリティ・データに対応するディスク・ドライブにおいてコマンド・キューイングが使用可能となり、高性能オペレーションが確保される。

【0019】

【実施例】図1を参照すると、データ処理システム11が示される。データ処理システム11はシステム中央処理ユニット(CPU)13、システム・メモリ15、大容量記憶制御装置17、及び通信リンク19を含み、通信リンク19はCPU13、システム・メモリ15及び大容量記憶制御装置17をリンクし、それらの間のデータ及びコマンドの交換を可能とする。通信リンク19はシステム・バスまたはあるタイプのネットワークを表す。

【0020】大容量記憶制御装置17はデータ・ブロックをRAID21にストライプし、またRAID21からデータ・ブロックを回復する機能を提供する。本発明はレベル4またはレベル5モードのRAID21の使用に適応する。大容量記憶制御装置17は通信リンク19とのインタフェース23を含む。インタフェース23はシステム・バス(例えばマイクロチャネル、EISAなど)またはSCSI接続、或いは通信リンク19へのネットワーク・アダプタである。ローカル・プロセッサ25とメモリ・バッファ27との間には、排他的論理和プロセッサ29が接続される。排他的論理和プロセッサ29は、RAID21において冗長情報として使用されるパリティ・データをデータ・ストライプに渡って生成し、性能を改良するために使用される。排他的論理和プロセッサ29はまたパリティ情報の更新を提供する。

【0021】様々なタイプのデータがメモリ・バッファ27上に記憶され、そうしたデータにはパリティ・キャッシュ・テーブル31及びパリティ・データ・ブロック33が含まれる。ユーザ・データ・ブロック35は、特にエラー回復のために遷移してメモリ・バッファ27を通過する。ローカル・プロセッサ25は通信リンク19からインタフェース23を介して受信されるデータを取得し、RAID21内の複数の直接アクセス記憶装置間でデータをストライプするように編成する。ローカル・プロセッサ25はまた、メモリ・バッファ27を用いてRAID21からデータを回復及び再編成し、インタフェース23を介してデータ処理システム11の計算ユニットに提供する。

【0022】ローカル・プロセッサ25はまた、メモリ・バッファ27内のパリティ・データ・ブロック33へのアクセスを直列化する処理を実施するために使用される。データは大容量記憶制御装置17から複数の入出力制御装置37A乃至37Dを介して、RAID21に渡される。各入出力制御装置37A乃至37Dはそれぞれローカル・バッファ39A乃至39Dをアクセスし、直接アクセス記憶装置またはディスク・ドライブ41A乃至41Dをそれぞれ制御する。各ローカル・バッファ39A乃至39D内には、それぞれコマンド・キュー43A乃至43Dが存在する。入出力制御装置37A乃至37Dは、コマンド・キュー43A乃至43D内に待機されるコマンドの実行順序を特定の制約内において最適化する。

【0023】コンピュータ・ベースのデータ処理システム11のシステム・メモリ15はオペレーティング・システム45を含み、これには大容量記憶制御装置17により実現される論理記憶装置のためのデバイス・ドライバを含む複数のデバイス・ドライバが含まれる。システム・メモリ15には更に、CPU13により大容量記憶制御装置17に送信されるコマンドのためのデバイス・ドライバ・キューが含まれる。

【0024】図2はディスク・ドライブ120の上面図であり、本発明を理解する上で有用なディスク・ドライブ・オペレーションの態様を表す。ディスク142の磁気表面140上の複数のトラック(1乃至N+1)の1つに対するアーム123によるスライダ126の位置決めが、トラック内のセクタ(セクタ146及び148など)からデータを読み出したり書込むために実行される。各トラックはディスク142の中心から放射状に広がる複数のトラック・サーボ・フィールド144により、セグメントまたはセクタに分割される。トラック・サーボ・フィールドは、回転アーム123の端部のスライダ126の移動に一致するように湾曲される。ディスク回転速度が一定の場合(すなわち一定角速度または"CAV"(constant angular velocity))、スライダ126に実装されるトランスジューサがトラック・サーボ・フィールド144に厳密な時間間隔で遭遇する。トラッキング情報が既知のように、トラック・サーボ・フィールド144から導出される。スライダ126はサーボ・フィールドの間を盲目的に浮上する。ユーザ・データまたはパリティ・データがディスク・ドライブ120上で更新される時、旧データが所定の位置から読出され、置換データが常にそこに書込まれる。例えばセクタ146がどのようなタイプのデータを含んでいようと、これは最初の回転において読出され、更新データが第2のまたは以降の回転においてセクタに書込まれる。セクタに対する読出しと書込みとの間に少なくとも1回転に対応する遅延が生じる。従来技術では、この間にパリティ・データを記憶するディスクへのアクセスが許可され

ず、ドライブ性能に影響を与えた。

【0025】図3は4つのディスク・ドライブ間における、ユーザ・データ及びパリティ・データ・ブロックのストライピングを示す。RAIDレベル4及びRAIDレベル5(左対称)の両方が表される。RAIDレベル4では、4つのストライプが示され、第1のストライプはドライブ0乃至3上にそれぞれデータ・ブロックD0、D1、D2及びP0を含み、第2のストライプはドライブ0乃至3上にそれぞれデータ・ブロック3乃至5及びパリティ・ブロック1を含み、第3のストライプはドライブ0乃至3上にそれぞれデータ・ブロック6乃至8及びパリティ・ブロック2を含み、第4のストライプはドライブ0乃至3上にそれぞれデータ・ブロック9乃至11及びパリティ・ブロック3を含む。RAIDレベル4では、第3のドライブはチェック・ディスクとして知られる。

【0026】RAIDレベル5では、任意のドライブがパリティ・データを含むために、チェック・ディスクを含まない。RAIDレベル5の一例では、RAIDレベル4と全く同じメンバシップを有するストライプを有する。しかしながら、第1のストライプ以降ではデータ分布は異なる。第2のストライプでは、パリティ・ブロックがドライブ2に移行し、第3のストライプでは、パリティ・ブロックがドライブ1に移行し、第4のストライプでは、パリティ・ブロックがドライブ0に移行する。第2のストライプでは、データ・ブロックが左回転され、端のブロックがドライブ3に移動する。第3のストライプでは、ブロックの平均変位は左方向に2ドライブとなり、4番目のストライプでは左方向に3ドライブとなる。RAIDレベル5は、チェック・ディスクへのアクセスの競合を回避するように設計される。

【0027】図4は、ドライブ1、3及び4のコマンド・キュー及びデバイス・ドライバ・キュー上でのコマンドのキューイングを表すタイミング図である。図4で考慮される例は、RAIDレベル4またはRAIDレベル5に適用され、ホストからの書き込み更新コマンド、すなわちドライブ0上のブロックD0を最初に更新し、次にドライブ2上のブロックD2を更新するように指示するコマンドに関連する。これらの更新の各々は同一のストライプに対応するパリティの更新を要求し、これはどちらの場合にもパリティ・ブロックP0に相当する。従って、P0へのアクセスの競合が発生する。更新のための各ホスト要求は更新を構成する付加データを含むか、データを検索するためのシステム・メモリ15内の位置を識別する。時刻T1において、データ・ブロック0にデータを書込むためのコマンドがデバイス・ドライバ・キュー47に存在する。時刻T2において、ブロック0内のデータを読み出し、付加データをブロック0に書込むためのリンク対コマンドがドライブ0のコマンド・キューに現れる。また時刻T2では、パリティ・ブロック0を

読出するための命令がドライブ3のキューに配置される。これはパリティ・ブロック0のデータがメモリ・バッファ27に保持されていることを示すキャッシュ・ヒットが発生しない場合に限り、発生する。

【0028】時刻T3において、データ・ブロックD2のデータを読出すコマンドがデバイス・ドライバ・キュー47に配置される。その後時刻T4において、データ・ブロックD2を読出すコマンドから導出されるリンク・コマンドが、ドライブ2のコマンド・キューに現れる。これにはデータ・ブロックD2の読出し及びデータ・ブロックD2の書き込みが含まれる。また時刻T4では、パリティ・ブロックP0を読出す命令がドライブ3のコマンド・キューを通じて進行するように示される。T4とT6との間のある時刻において、パリティ・ブロックP0がメモリ・バッファ27に読出される。また時刻T6より先にデータ・ブロックD0を読出す命令が実行される。時刻T6において、更新データをデータ・ブロックD0に書き込むコマンドが、ドライブ0のコマンド・キューの出力端に達するように示される。このコマンドは時刻T12より以前に実行される。時刻T7において、パリティ・ブロックP0のデータを置換する書き込み命令が、ドライブ3のコマンド・キューに配置される。このコマンドの進行は時間間隔T8を経過して継続される。時刻T9において、ユーザ・データ・ブロックD2の読出し命令がドライブ2のコマンド・キューから実行され、更新データをデータ・ブロックD2に書き込む命令がコマンド・キューの出力端に達する。実行はT9以降に発生する。時刻T10において、パリティ・ブロックP0への書き込み命令がドライブ3のコマンド・キューの入力端に配置される。時刻T10の時点で、パリティ・ブロックP0に対する2つの書き込み命令がドライブ3のコマンド・キュー内に同時に存在する。時間間隔T11及びT12では、パリティ・ブロックP0に対する2番目の書き込み命令がドライブ3のコマンド・キュー内において進行する。

【0029】図5はパリティ・キャッシュ・テーブル31の構造を表す。パリティ・キャッシュ・テーブルは、メモリ・バッファ27に存在するパリティ・データ・ブロックに関連する最大K個のエントリを含む。別の実施例では、パリティ・キャッシュ・テーブル31及びパリティ・データ・ブロック33が、システム・メモリ15に記憶される。当業者には、メモリ・バッファ27内のデータへのアクセスが、システム・メモリ15へのアクセスよりも高速であることが望ましいことが理解されよう。いずれにしても、これらのいずれのアクセスもディスク・ドライブからの回復よりは高速である。データ構造を記憶するためのシステム・メモリ15の使用により、通信リンク19上のトラフィックは増加する。

【0030】パリティ・キャッシュ・テーブル31内の各エントリは、RAIDレベル4システムのドライブ・

ユニット41D、またはRAIDレベル5システムのドライブ41A乃至41Dの1つにおける論理ブロック・アドレス範囲に対応する論理ブロック・アドレス範囲を含む。アドレス範囲フィールド70は、ドライブ・ユニット上の開始論理ブロック・アドレス及びパリティ・ブロックの終りを含む。エントリには更に、エントリに関連するパリティ・データへの最も最近の使用またはアクセスのタイム・スタンプ71が含まれる。NWW (number of writeswaiting) フィールド73は、実行を待機中の書き込みの数を示すカウンタである。フィールド75 (L/F) はデータへのアクセスをロックするためのフラグであり、メモリ・バッファ27内のパリティ・データの直列化の変更を提供する。最後に、データ・キャッシュ・テーブル31内のエントリは、パリティ・データ・ブロックを見出すためのバッファ内における位置を示す。データ・キャッシュ・テーブル31及びパリティ・データ・ブロック33は、電力損失を防ぐために、不揮発メモリ内に保持されてもよい。

【0031】図6は、ホスト処理ユニットからの更新コマンド、すなわちRAID21へのデータの読出し—変更—書き込みを要求するコマンドの受信において入力される処理の論理流れ図である。付加データの識別を含む更新コマンドの受信に際し、適切なストライプのデータ及びパリティ・データに対応する論理ブロック・アドレス範囲がステップ201で計算される。プログラムはユーザ・データの更新及びパリティ・データの更新のための処理を両立するが、これは実質的には非同期処理であり、ドライブが非同期に動作する。ユーザ・データ処理ではステップ203へ移行し、適切なデータ・ドライブ・コマンド・キューがフルかどうかを判断する。フルの場合、処理はエントリがキュー内で開くまで待機する。ユーザ・データ・ドライブのコマンド・キューの空間が使用可能になると、分岐せずにステップ203からステップ205に移行する。ステップ205で、1) ステップ201で決定された論理ブロック・アドレスから旧データを読出し、2) 付加データを同一位置に書き込むために、リンク・コマンドがデータ・ドライブ・コマンド・キューに送信される。処理は次にステップ207に移行し、付加データと置換パリティ・データの両方が書き込まれたことが示されるまで、待機オペレーションを実行する。ステップ207は実質的に2つの処理の再同期化を提供する。ステップ207の後、ステップ209が実行され、完全なステータスをホストに送信する。処理はこの時点で終了する。

【0032】パリティ・データ処理ではステップ211が実行され、パリティ・ドライブ・キューがフルかどうか判断される。フルの場合、パリティ・データを保持するドライブのコマンド・キューに、コマンドのための空間が使用可能になるまで、待機が実行される。空間が使用可能となると分岐せずにステップ213に移行す

る。ステップ213では、ステップ201でパリティ・データに対して計算されたパリティ・キャッシュ・テーブルの論理ブロック・アドレス範囲に、エントリが存在するかどうか判断される。パリティ・キャッシュ・ヒットが発生すると、ステップ213からステップ215へ分岐する。ステップ215では、パリティ・データ・ブロックのステータスがフリーかどうか判断される。フリーでない場合、ステータス・フラグがフリーになるまで待機が実行される。次にステップ215からステップ217に分岐し、ステータス・フラグをロックにセットし、待機中の書込み数を増分し、パリティ・キャッシュ・エントリ内のタイム・スタンプを置換する。ステップ217に続きステップ219が実行され、新たなまたは置換パリティを計算するために要求される旧データが、ステップ205から使用可能かどうかを判断する。処理はステップ219で旧ユーザ・データが使用可能となるまで待機する。データが使用可能になると、ステップ221に分岐し、新たなパリティ・データを計算し、キャッシュ内のパリティ・ブロックを更新する。ステップ223でステータス・フラグがフリーにセットされ、ステップ225で新たなパリティ・データの書込みコマンドが適切なディスク・ドライブに送信され、その入出力制御装置に対応するコマンド・キューに待機される。ステップ227は、入出力制御装置がパリティ書込みオペレーションの完了を示すまで待機サイクルを提供する。書込みが完了するとステップ229が実行され、待機中の書込み数が1減分される。処理は次にステップ207に移行する。

【0033】ステップ213でパリティ・キャッシュ・ヒットが発生しなかったと判断されると、分岐せずにステップ231に移行する。ステップ231では、フリー・パリティ・キャッシュ・エントリが使用可能かどうか判断される。使用可能な場合、ステップ233で、フリー・パリティ・キャッシュ・エントリを使用し、新たなパリティ・キャッシュ・エントリが作成される。使用可能でない場合には、ステップ235でパリティ・キャッシュ・エントリがテーブルから解放される。これは待機中の書込み数が0である最低使用頻度(LRU)パリティ・キャッシュ・エントリを見い出すことにより実行される。エントリが探し出されると、ステップ233で新たなパリティ・エントリが、解放されたエントリから作成される。次にステップ237でステータス・フラグがロックにセットされ、待機中の書込み数が1にセットされる。次にステップ239で旧パリティがパリティ・ドライブから読出される。これはステップ241で読出しオペレーションの完了として示されるように、コマンドの通常のキューイング及び待機を含むものと理解される。パリティの読出しが完了し、その読出されたばかりのパリティのバッファ27における記憶位置がキャッシュ・エントリに含まれると、ステップ241からステッ

プ219に分岐し、処理が上述のように継続される。

【0034】入出力制御装置レベルにおけるタグ・コマンド・キューイングによる性能の改良は、こうしたシステムが本発明を使用するRAIDレベル4またはレベル5大容量記憶システムに適用されると失われることはない。キャッシュ・エントリ・テーブルのデータ構造が、電源障害による損失から保護するために不揮発RAMにより実現されてもよい。

【0035】本発明は特定の実施例に関連して述べられてきたが、当業者には本発明の精神及び範囲から逸脱することなく、形態及び詳細に関する様々な変更が可能であることが理解されよう。

【0036】まとめとして、本発明の構成に関して以下の事項を開示する。

【0037】(1) 耐障害論理大容量記憶装置内の物理記憶装置にユーザ・データ及び冗長データを記憶する方法であって、ホストからの付加データを伴う書込みコマンドの受信に応答して、第1の物理記憶装置上における付加データの論理ブロック・アドレスと、第2の物理記憶装置上における冗長データのパリティ・アドレスとを決定するステップと、前記第1の記憶装置に対して、論理ブロック・アドレスの旧データを読出し、論理ブロック・アドレスに付加データを書込むように指令するステップと、前記第2の記憶装置上の冗長データがメモリ・バッファ内に存在するかどうかを判断するステップと、冗長データが前記メモリ・バッファ内に存在しない場合、前記第2の記憶装置から前記メモリ・バッファに冗長データを読出すステップと、冗長データ、付加データ及び旧データから新たな冗長データを計算するステップと、前記第2の物理記憶装置に新たな冗長データを論理ブロック・アドレスに記憶するように指令するステップと、を含む記憶方法。

(2) 新たな冗長データを計算するステップが、前記メモリ・バッファ内の冗長データを新たな冗長データにより置換するステップを含む、前記(1)記載の記憶方法。

(3) ディスク・ドライブの冗長アレイ内のパリティ・ドライブ上にコマンドをキューイングする方法であって、ホストからの付加データを伴う書込みコマンドの受信に応答して、データ・ドライブ上における付加データの論理ブロック・アドレスと、前記パリティ・ドライブ上におけるパリティ・データの論理ブロック・アドレスとを決定するステップと、パリティ・キャッシュ・エントリがパリティ・データの論理ブロック・アドレスに一致するかどうかを判断するステップと、前記一致の肯定判定に応答して、パリティ・キャッシュ・エントリにより提供されるバッファ・ロケーションのパリティ・データと、前記データ・ドライブから読出されるデータと付加データとから、置換パリティ・データを計算するステップと、置換パリティ・データを前記バッファ・ロケー

ションに配置するステップと、置換パリティ・データを
書込むためのコマンドを、パリティ・データの論理ブ
ロック・アドレスに対応する入出力制御装置に送信するス
テップと、を含むキューイング方法。

(4) 前記一致の否定判定に応答して、パリティ・デー
タを前記バッファ・ロケーションに読出すためのコマン
ドを、前記パリティ・ドライブの論理ブロック・アドレ
スに対応する入出力制御装置に送信するステップと、前
記パリティ・ドライブの論理ブロック・アドレスのパリ
ティ・キャッシュ・エントリを生成するステップと、を
含む、前記(3)記載のキューイング方法。

(5) パリティ・キャッシュ・エントリが前記パリティ
・ドライブの論理ブロック・アドレスと、ドライブ識別
と、タイム・スタンプと、ロック/非ロック・フラグ
と、待機中の書込みの数を示す書込み待機カウンタと、
メモリ・バッファ内のパリティ・データの位置を示す位
置フィールドとを含む、前記(4)記載のキューイング
方法。

(6) 前記データ・ドライブの論理ブロック・アドレス
の決定に応答して、データを読出し次に付加データを書
込むためのリンク・コマンドを、論理ブロック・アドレ
スをカバーする入出力制御装置に送信するステップ、を
含む、前記(5)記載のキューイング方法。

(7) 計算ステップがリンク・コマンドの送信ステップ
の読出しオペレーションの後に発生する、前記(6)記
載のキューイング方法。

(8) 前記パリティ・キャッシュ・エントリの生成ステ
ップが、パリティ・キャッシュ・テーブルが未使用のパ
リティ・キャッシュ・エントリを有するかどうかを判断
するステップと、肯定判定に応答して、未使用のパリ
ティ・キャッシュ・エントリを使用するステップと、否定
判定に応答して、待機中の書込みを有さないパリティ
・キャッシュ・エントリの中から、最低使用頻度のパリ
ティ・キャッシュ・エントリを探し出し、探し出されたパ
リティ・キャッシュ・エントリを使用するステップと、
を含む、前記(5)記載のキューイング方法。

(9) 置換パリティ・データ書込みのためのコマンドの
送信ステップが、書込み待機カウンタを増分するステ
ップを含む、前記(5)記載のキューイング方法。

(10) 置換パリティ・データの生成のために、前記バ
ッファ・ロケーション内のパリティ・データへのアクセ
スを直列化するステップ、を含む、前記(9)記載のキ
ューイング方法。

(11) 置換パリティ・データが前記パリティ・ドライ
ブの論理ブロック・アドレスに書込まれた後、パリティ
・キャッシュ内の書込み待機カウンタを減分するステ
ップ、を含む、前記(10)記載のキューイング方法。

(12) 前記パリティ・ドライブがチェック・ディスク
である、前記(11)記載のキューイング方法。

(13) 前記パリティ・ドライブがディスク・ドライブ

の冗長アレイの任意の1つに相当する、前記(11)記
載のキューイング方法。

(14) ホスト・データ処理システムからの付加データ
を伴う書込み命令の受信に応答して、データ・ドライ
ブ上における付加データの論理ブロック・アドレスと、パ
リティ・ドライブ上における冗長データの論理ブロック
・アドレスとを決定する手段と、複数のパリティ・キャ
ッシュ・エントリを有するパリティ・キャッシュ・テー
ブルと、パリティ・キャッシュ・エントリ及び前記パ
リティ・ドライブ上の論理ブロック・アドレスに対応する
パリティ・データを記憶するバッファと、パリティ・キ
ャッシュ・エントリが発見手段により見い出される前記
パリティ・ドライブの論理ブロック・アドレスに一致す
るかどうかを判断する手段と、前記一致の肯定判定に
応答して、パリティ・キャッシュ・エントリにより提供さ
れるバッファ・ロケーションのパリティ・データと、前
記データ・ドライブの論理ブロック・アドレスから読出
されるデータと、付加データとから、置換パリティ・デ
ータを計算する手段と、置換パリティ・データを前記バ
ッファ・ロケーションに配置する手段と、置換パリティ
・データを書込むためのコマンドを、パリティ・ドライ
ブの論理ブロック・アドレスに対応する入出力制御装置
に送信する手段と、を含む、ディスク・ドライブの冗長
アレイ。

(15) 前記一致の否定判定に応答して、パリティ・デ
ータを前記バッファ・ロケーションに読出すためのコマ
ンドを前記パリティ・ドライブの論理ブロック・アドレ
スに対応する入出力制御装置に送信する手段と、前記パ
リティ・ドライブの論理ブロック・アドレスから読出さ
れるパリティ・データに対応するパリティ・キャッシュ
・エントリを生成する手段と、を含む、前記(14)記
載のディスク・ドライブの冗長アレイ。

(16) パリティ・キャッシュ・エントリが前記パリ
ティ・データの論理ブロック・アドレスと、タイム・スタ
ンプと、ロック/非ロック・フラグと、待機中の書込み
の数を示す書込み待機カウンタと、ドライブ識別と、メ
モリ・バッファ内のパリティ・データの位置フィールド
とを含む、前記(15)記載のディスク・ドライブの冗
長アレイ。

(17) 前記データ・ドライブの論理ブロック・アドレ
スの決定に応答して、データを読出し次に付加データ
を書込むためのリンク・コマンドを、前記データ・ドライ
ブの論理ブロック・アドレスに対応する入出力制御装置
に送信する手段、を含む、前記(16)記載のディスク
・ドライブの冗長アレイ。

(18) パリティ・キャッシュ・エントリの生成手段
が、パリティ・キャッシュ・テーブルが未使用のパリ
ティ・キャッシュ・エントリを有するかどうかを判断する
手段と、肯定判定に応答して、未使用のパリティ・キャ
ッシュ・エントリを使用する手段と、否定判定に応答し

て、待機中の書込みを有さない最低使用頻度のパリティ・キャッシュ・エントリを探し出し、探し出されたパリティ・キャッシュ・エントリを使用する手段と、を含む、前記(16)記載のディスク・ドライブの冗長アレイ。

(19) 置換パリティ・データ書込みのためのコマンドの送信手段が、書込み待機カウンタを増分する手段を含む、前記(16)記載のディスク・ドライブの冗長アレイ。

(20) 置換パリティ・データの生成のために、前記バッファ・ロケーション内のパリティ・データへのアクセスを直列化する手段、を含む、前記(19)記載のディスク・ドライブの冗長アレイ。

(21) 置換パリティ・データが前記パリティ・ドライブの論理ブロック・アドレスに書込まれた後、前記パリティ・ドライブの論理ブロック・アドレスに対応するパリティ・キャッシュ内の書込み待機カウンタを減分する手段、を含む、前記(20)記載のディスク・ドライブの冗長アレイ。

(22) 前記パリティ・ドライブがチェック・ディスクである、前記(21)記載のディスク・ドライブの冗長アレイ。

(23) 前記パリティ・ドライブがディスク・ドライブの冗長アレイの任意の1つに相当する、前記(21)記載のディスク・ドライブの冗長アレイ。

(24) 論理装置として動作する物理記憶装置のアレイと、物理記憶装置のアレイ上の所定の位置に記憶される付加データを伴う更新コマンドを提供するコマンド発生器と、ユーザ・データのストライプに対応する選択パリティ・データ・ブロックのコピーを記憶するキャッシュと、ユーザ・データ・ブロック及びユーザ・データ・ブロックのストライプに渡って生成されるパリティ・データ・ブロックを物理記憶装置のアレイ間でストライプし、物理記憶装置位置によりキャッシュ内に記憶される選択パリティ・データ・ブロックの1個以上のコピーを識別するパリティ・キャッシュ・テーブルを保守するローカル・プロセッサと、ローカル・プロセッサによりそれぞれの入出力制御装置に送信されるコマンドをキューイングする、物理記憶装置の物理アレイの各々に関連する入出力制御装置と、を含むデータ処理システム。

(25) パリティ・キャッシュ・エントリが物理記憶装置の論理ブロック・アドレスと、物理記憶装置の識別と、タイム・スタンプと、ロック/非ロック・フラグと、待機中の書込みの数を示す書込み待機カウンタと、キャッシュ内のパリティ・データ・ブロックの位置・フィールドとを含む、前記(24)記載のデータ処理システム。

(26) ローカル・プロセッサが、キャッシュ内の選択パリティ・データ・ブロックの各コピーへのアクセスを制御するシリアルライザ、を含む、前記(25)記載のデ

ータ処理システム。

(27) キャッシュ内にパリティ・データ・ブロックが存在しないストライプ内の物理アドレスを更新するコマンドに応答して、前記ストライプに対応するパリティ・データ・ブロックをキャッシュに読出し、パリティ・キャッシュ・テーブルに前記パリティ・データ・ブロックのエントリを生成するパリティ・キャッシュ・テーブル・エントリ発生器、を含む、前記(26)記載のデータ処理システム。

10 (28) パリティ・データが単一の物理記憶装置上に配置される、前記(27)記載のデータ処理システム。

(29) パリティ・データが全ての物理記憶装置上に配置される、前記(27)記載のデータ処理システム。

【0038】

【発明の効果】以上説明したように、本発明によれば、集散的に1個以上の論理大容量記憶装置として機能する大容量記憶装置のセットが提供され、RAIDレベル4及び5システムにおけるパリティ・ドライブ上で、コマンドをキューイングするシステム及び方法が提供され

20 る。

【図面の簡単な説明】

【図1】RAIDを含むデータ処理システムのハイレベル・ブロック図である。

【図2】ディスク・ドライブの上面図である。

【図3】RAIDレベル4及び5システムにおけるデータ・ストライピングを表す図である。

【図4】RAID4または5システムにおいてストライプされるデータを変更するコマンド・シーケンスのタイミング図である。

30 【図5】本発明をサポートするために使用されるパリティ・キャッシュ・テーブルのデータ構造を表す図である。

【図6】本発明を実施するために、RAIDシステムの記憶制御装置上で実行される処理のハイレベル論理流れ図である。

【符号の説明】

11 データ処理システム

13 中央処理ユニット(CPU)

15 システム・メモリ

40 17 大容量記憶制御装置

19 通信リンク

21 RAID

23 インタフェース

25 ローカル・プロセッサ

27 メモリ・バッファ

29 排他的論理和プロセッサ

31 パリティ・キャッシュ・テーブル

33 パリティ・データ・ブロック

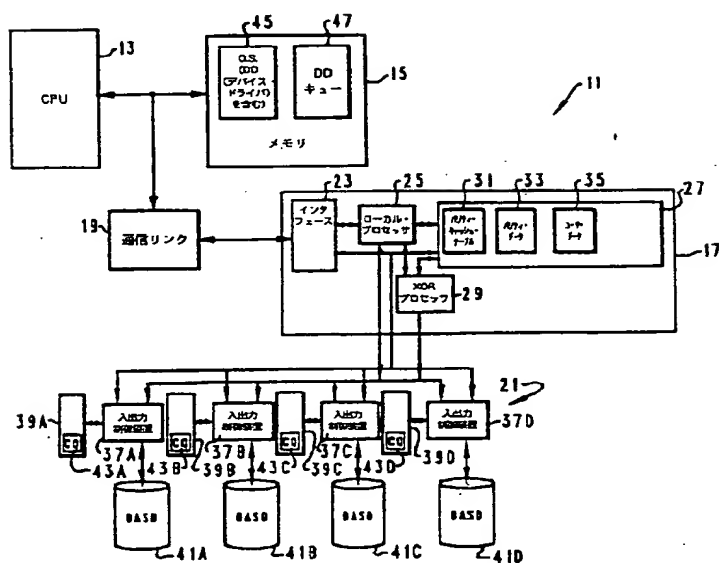
35 ユーザ・データ・ブロック

50 45 オペレーティング・システム

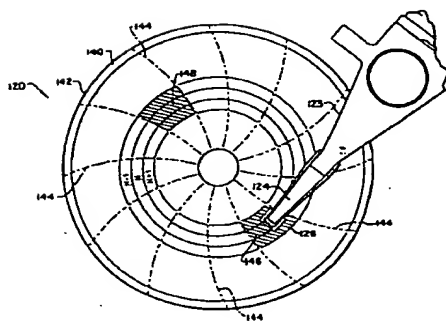
21
 47 デバイス・ドライバ・キュー
 70 アドレス範囲フィールド
 71 タイム・スタンプ
 73 NWWフィールド
 75 フィールド
 120 ディスク・ドライバ

22
 123 アーム
 126 スライダ
 140 磁気表面
 142 ディスク
 144 トラック・サーボ・フィールド
 146、148 セクタ

【図1】



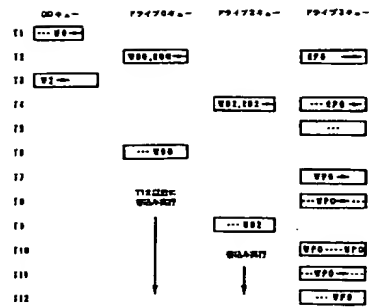
【図2】



【図3】

RAID LEVEL 0				RAID LEVEL 5 (CENTRO)			
0000	0001	0002	0003	0000	0001	0002	0003
00	01	02	03	00	01	02	03
04	05	06	07	04	05	06	07
08	09	10	11	08	09	10	11

【図4】



【図5】

アドレス範囲	70	71	73,75	77	位置
...
...
...
...
...
...
...

【図6】

